TECHNICAL REPORT SERIES

THE ECONOMICS SERIES

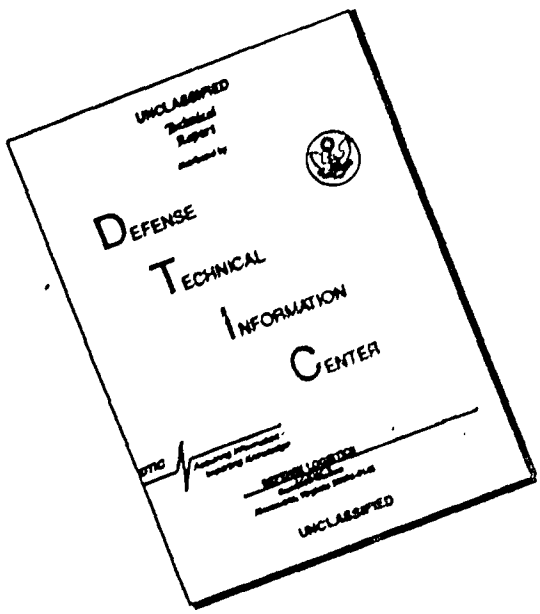
INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES

FOURTH FLOOR ENCINA HALL

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

EFFICIENT AND COMPETITIVE RATIONING

by

Robert Wilson

Technical Report No. 528

May 1988

A REPORT OF THE
CENTER FOR RESEARCH ON ORGANIZATION EFFICIENCY
STANFORD UNIVERSITY
CONTRACT N00014-~~K-0216~~ 86-K-0216
United States Office of Naval Research,
the National Science Foundation SES 86-05666,
and the Electric Power Research Institute

THE ECONOMICS SERIES

INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES
Fourth Floor, Encina Hall
Stanford University
Stanford, California 94305

EFFICIENT AND COMPETITIVE RATIONING*

by

Robert Wilson†

INTRODUCTION

In a spot market prices are varied continually to balance supply and demand. Many of the standard durable and storable commodities are traded in spot markets. However, several important industries produce the non-storable services of capital and labor. Standard examples are the capital-intensive industries subject to peak loads, such as electric power, telecommunications, transport, and hotels. Others with similar characteristics include the make-to-order industries (e.g., paper, metalworking) and the various industries of the service sector that essentially rent capacity or servers to customers. In these industries also, one might expect spot markets to be used to achieve efficient allocation of scarce supplies. Indeed, theorists have often argued that ideally prices should be varied continually to keep demand within capacity while ensuring an efficient allocation of scarce supplies among customers. The theory of spot pricing and its variants such as peak load pricing represent this view. Vickrey (1971) argues that there are substantial efficiency gains to be realized from wider use of spot pricing, and a

---

*Presented as the Fisher-Schultz Lecture at the 1986 European Meetings of the Econometric Society in Budapest.

† The work reported here includes joint research with Hung-po Chao, Shmuel Oren, and Stephen Smith. Research support from the Office of Naval Research, the National Science Foundation (SES8605666), and the Electric Power Research Institute is gratefully acknowledged.

particular application to electric power is developed by Bohn et al. (1984) and Caramanis et al. (1982).

Actually, spot pricing is rarely used to ration supplies in these industries, even the standard examples cited in theoretical studies. The apparent explanations are technological limitations and pervasive transactions costs. For example, it may be difficult or expensive to inform a customer continually about prices and to monitor the time pattern of purchases, even if one were to know the right prices to sustain capacity utilization and to prevent congestion. As a result, it is commonplace to see fixed prices that lead to idle capacity in slack periods, and random or arbitrary allocation of scarce supplies in peak periods. These practices impose costs on customers forced to queue for service and on those denied service by price or quantity rationing. Allocative inefficiencies result when customers' preferences differ and their service orders do not conform to the ordering of their preferences.

Spot prices and fixed prices are not the only possibilities. In a few cases, state enterprises (regulated public utilities in the United States) and private firms use contingent forward contracts to improve allocative efficiency. Some of these contracts, such as reservations that ensure an allotment of capacity to the user, are unconditional forward sales. Other contracts, such as interruptible service contracts for electric power and natural gas, condition service on particular events.

The contracts that interest us here are called priority service contracts. The salient feature of such contracts is that they specify each customer's priority in obtaining service. That is, they specify the rank order in which a customer is served out of the available supply, until all

customers are served or supply is exhausted. Such contracts essentially establish queues for customers.

In Section 1 the author provides some background about priority service. In Section 2 we formulate a basic model and offer several illustrations. We also described two main examples that motivate the theoretical development. In Section 3 we derive some key results that show how the prices of priority service contracts are designed to induce customers to self-select efficient service orders. In Section 4 the author discusses various ways that state enterprises can organize markets that implement priority service efficiently. In Section 5 we study the operation of competitive markets for priority service. We conclude in Section 6 concludes with some summary remarks.

The themes are We emphasize two themes. One is that a state enterprise can promote substantial efficiency gains by substituting priority service for absent spot markets. The other is that oligopolistic firms may have insufficient incentives to offer efficient product diversity; consequently, allocative efficiency depends on entry of numerous firms. Even so, dispersal of supplies among many firms can prevent productive efficiency when there are advantages from pooling supplies. Keywords: Scarce Supplies

1. BACKGROUND

I recently studied the electric power industry in the United States, so I use it to illustrate.

The demand for power is subject to long trends and cycles that vary with economic conditions, to strong seasonal and daily cycles, and to fairly sudden surges: the annual peak may occur briefly on a hot summer afternoon

or a cold winter night. On the supply side, reservoirs are subject to seasonal and multi-year cycles, generators need to be shut down for maintenance, and equipment can fail suddenly. Power is essentially non-storable except as potential energy, so shortages must be matched quickly by interruptions; often the warning period is too short to enable a spot market to operate. Until recently, supply shortages were mostly allocated indiscriminately by curtailing service to entire districts.

The development of inexpensive micro-electronic devices, however, makes it possible to interrupt customers selectively by using radio or on line-frequency signals to activate circuit breakers in their meters. This allows priority service contracts with the following form. In addition to the direct charge for energy, a customer pays a premium depending on the priority class he selects. In the simplest case, two priorities are offered and the customer assigns a base portion of his load to high priority and the residual to low priority.[1] When interruptions are necessary, no customer's high-priority base load is interrupted unless there is still a shortage after <u>all</u> customers' low-priority loads are interrupted. The premia can be refunded as dividends to customers or shareholders, or they can finance additions to capacity.

Continuing advances in metering devices and control procedures enable a richer array of options. In addition to selecting among a larger number of priority classes, customers can elect one priority for interruption and another for resumption of service after an interruption. A customer who incurs shutdown and restart costs when interrupted prefers a higher priority for interruption than resumption.

In addition to the advantages for customers, these contracts enable the

enterprise to substitute low-priority interruptible loads having relatively low value to customers for expensive additions to capacity that would otherwise be required to sustain reliable service to valuable end uses. That is, like reserve capacity, contracts for low-priority service provide an inventory of 'supply' to meet shortfalls and thereby protect the higher reliability expected from high-priority contracts.[2]

The Role of Priority Service

The example from the power industry illustrates several points. First, compared to random rationing of scarce supplies, priority service provides efficiency gains by serving customers in the order that conforms to the costs they incur from interruption. The source of these efficiency gains is that, without spot markets to ration supplies efficiently, customers suffer pecuniary externalities in the form of unpriced congestion. Priority service alleviates these effects by offering a forward market for service orders.[3]

From a customer's viewpoint, the perceived advantage of priority service is the increased product differentiation: each customer can select from the schedule of priorities and their associated prices. Whenever customers differ in their interruption costs or service values, the product differentiation offered in a menu of priority service options promotes more efficient allocation of supplies.

The variety of service conditions that can be offered is constrained primarily by the technology and costs of monitoring, metering, and control. The prices are affected substantially by the fact that each customer's selection is based on private information about his preferences. That is,

customers' self-selection of their options imposes constraints on the relative prices of any two options if these prices are to implement efficient service orders.

The menu is further constrained by the technology of supply. The quality attributes of the various options (e.g., reliability or speed of service) are determined jointly by the probability distributions of demands and supplies and by the numbers of customers selecting each option. Customers' perceptions of quality attributes must be 'rational expectations' if they are to match correctly the qualities that can be delivered by the technology available.

As one expects, the premium that a customer pays for priority service is, in the simplest models, simply the mathematical expectation of the spot prices that would be paid for equivalent service if spot markets were operating. Compared to spot markets, therefore, priority service is an innovative form of contracting that reduces the cost of market organization. However, there are unequivocal advantages from priority service even in the absence of transaction costs. If premia are refunded <u>equally</u> to customers as dividends, then priority service is Pareto superior to random rationing. That is, <u>every</u> customer benefits from the adoption of priority service and the seller's revenue is not reduced.

As noted above, one way to interpret priority service is that it is a kind of product differentiation. The quality attributes of the conditions of delivery are differentiated to allow customers with different preferences to select different qualities. It differs from ordinary product differentiation, however, in that the qualities obtained are endogenous: they depend on how many other customers select the same and higher

priorities. Different prices at retail stores offering generically identical products are a familiar example: one gets quicker service or more attention from servers by patronizing a higher-priced store having less demand. Priority service is possibly a major explanation for the dispersion of prices found in practice.

An alternative interpretation is that priority service is a precursor to insurance against the social risk of supply shortages. Actually, priority service only assures customers their rank orders of service. Efficient service orders assure that social risks are allocated among customers to minimize total losses. Thus, it is a preliminary to fully efficient insurance that includes compensation for losses from interruptions. If customers are risk averse then compensatory insurance can supplement priority service to obtain further efficiency gains.

Another benefit of priority service is purely informational. A public enterprise that offers a single service quality to all customers (e.g., it allocates shortages randomly) has no direct measure of customers' willingness to pay for capacity increments that improve the quality of service. This is a persistent problem in state enterprises (cf. Boiteux (1960) and Marchand (1974)), and in the case of electric power capacity in the United States it has been studied by Telson (1975). In contrast, customers' selections of priority service conditions reveal willingness to pay for quality improvements.

From a technical viewpoint, priority service is based on the supposition that customers' service orders determine allocative efficiency. This supposition is correct only when end uses have some nonconvexity (e.g., indivisibility or economy of scale) or informational effect in the

utilization of the service supplied. For example, customers may queue for access to servers because a server can not divide attention among several customers simultaneously without impairing productivity; thus there is a one-at-a-time rule. In the case of passenger transport, a vehicle serves a customer entirely or not at all. In principle, power could be rationed by reducing power proportionately to all customers, and actually this is feasible for end uses such as heating, and also air conditioning and pumping via 'cycling' (automatic periodic interruptions of service). In many applications, however, power is a productive input that enters in a fixed proportion or at a minimum efficient scale. Unexpected interruption or diminution of power therefore wastes the complementary factors of production. Because of these fundamental nonconvexities and informational effects underlying the application of priority service, parts of the subject lie outside the domain of standard microeconomic theory. An important feature is the incompleteness of the market for spot trading, due to the inability of firms and customers to communicate instantaneously. We shall see, nevertheless, that familiar methods can be adapted to solve most of the problems that arise.

On general grounds, the advantages of priority service derive from several key features of the situations in which it is used. One is that service must occasionally be rationed, queued, or otherwise differentiated in the quality attributes of the conditions of delivery. The second is that customers have diverse preferences so that there are efficiency gains from differentiation. The third, of course, is that spot markets are more difficult or expensive to operate than is a system that has customers select periodically among contingent forward contracts. If customers' preferences

are sufficiently persistent over time, then priority service is an efficient market organization.

Variants of priority service are evident in many forward contracts. For instance, the principle that the price of the forward contract is the expectation of the prices that would prevail in spot markets provides the simplest interpretation of prices for reserved accommodations with airlines and hotels, rates for the several classes of express mail, etc. The principle that service is offered contingent on adequate supply for the class of service selected is evident as well in the differing average load factors for first-class and tourist sections of airplanes, private and regular telephone lines, etc. Higher prices for quick or uncongested service partially explain the dispersion of prices among competing retail stores.

In other cases, the potential gains from priority service are unrealized. An important case is electric power, which is used in situations as diverse as hospital operating rooms and delicate production operations (high priority), and heating, air conditioning and agricultural pumping (low priority, partly because they are easily deferred for moderate durations). New metering and control technology makes electric power the prime candidate in which to introduce innovative forms of priority service. We therefore concentrate on implementations adapted to this industry.

Lastly we mention that the theory of priority service is in some ways dual to the theory of auctions. We develop the theory mainly for the case that aggregate demand is certain and supply is uncertain, although some extensions to the case of uncertain demand are made in Chao and Wilson (1987a). In contrast, the theory of auctions studies problems in which

aggregate demand is uncertain and supply is fixed; e.g., Harris and Raviv (1981). We do not attempt a unified theory here, but there is a prospect of a unified construction.

2. A MODEL AND EXAMPLES OF PRIORITY SERVICE

The model and examples presented in this section are simplified by several general assumptions. One is that the model is static: after contracting at an initial date, demand and supply occur at a single future date. Because the model allows aggregation over many future dates, this is not a significant restriction. A key assumption is that each customer is fully informed about his preferences; a weakening of this assumption is discussed in Section 4. For now, it rules out that at the time of contracting the customer is unsure about his later preferences; it also excludes random demand. Another assumption is that the seller knows the distribution of customers' preferences in the population. This is unrestrictive in practice because over time the distribution is revealed by customers' selections, provided the distribution is stable. Risk aversion is excluded except when we examine priority insurance in Section 3.4. The probability distribution of supply is assumed known to all participants. The firm observes the realized supply but the customers do not. And finally, regarding the seller's costs of production, we assume a constant marginal cost up to an inelastic quantity of supply, and for simplicity this marginal cost is normalized to zero; Chao and Wilson (1987a) address the case that marginal cost is increasing. That is, customers always pay the marginal cost of service and so their valuations of service are interpreted to be net of this cost.

Formulation

The ingredients of the formulation are as follows:

Customers' Preferences

Assume for simplicity that each customer demands a single unit of supply at any price not exceeding some privately known reservation value. This is unrestrictive in the main applications since separate units of demand can be treated separately.[4] Each customer's preferences are identified by a nonnegative number v. In the examples this number represents the value of one unit of service or the cost of an interruption. We call v the customer's type.

The population of customers is taken to be a continuum whose total measure is normalized to 1. Thus, if $H(v)$ represents the distribution of types in the population, then also $H(v)$ is the measure of customers with types not exceeding v and $\bar{H}(v) = 1-H(v)$ is the measure of types exceeding v. Note that $\bar{H}(0) = 1$, and by normalizing the scale in which the types are measured we can make $\bar{H}(1) = 0$; that is, the maximum type among the customers is $\bar{v} = 1$. Although it is not necessary, simplify by assuming that there are no large sets of customers of the same type - in the sense that H has a positive density and $\bar{H}$ has an inverse. In the examples, $\bar{H}$ and $\bar{H}^{-1}$ are the demand and inverse-demand functions.

The quality of service is also represented by a single number, say w. Absent income effects and risk aversion, therefore, a customer's net benefit has the form $u(v, w) - p$ if his type is v, the quality of service is w, and the price is p. Customers who forego service obtain the net benefit

0. Assume that u is increasing in each argument, and satisfies the usual self-selection condition that its cross-partial derivative exists and is positive: $u_{vw} > 0$. In the examples $u(v,w) = vw$ so that v measures the customer's valuation per unit of the quality w. This special form captures the relevant features of more general preferences.

Technology and Quality

Assume that customers impose identical requirements for service: customers differ only in their valuations of service quality, not in their service requirements. We interpret the technology in a reduced form that specifies the quality received by the customer in terms of the customer's service order. Thus, quality is a function $w(r)$ of the customer's service order r. It suffices to let the service order r be a number in the unit interval, with the interpretation that customers with smaller service orders have higher priority in obtaining service. In some cases we use a function $w(r, s)$ of both the service order and the supply s that is available. Interpreting s as the available supply per customer, we illustrate with examples having the special form $w(r, s) = w(r/s)$. Assume that w is a bounded, decreasing function of the service order r (i.e., increasing in priority), and an increasing function of the supply s. Adopt the normalization $u(0, w(1)) = 0$: this says that the lowest type is indifferent about receiving the last service order at marginal cost. Thus, the set of customers is identified with the set of those eligible for service in an efficient allocation.

If all customers are served in random order, then each customer receives the expected quality $\bar{w} = \int_0^1 w(r)dr$. Similarly, if only

$\bar{r} < 1$ customers are served in random order, then the expected quality is $\bar{w}(\bar{r}) = \int_0^{\bar{r}} w(r)dr/\bar{r}$ for the eligible customers. For instance, if $w(r) = 1 - r$ then $\bar{w}(\bar{r}) = 1 - \bar{r}/2$. If customers select priority classes indexed by service orders ρ, and $N(\rho)$ customers select service orders no higher than ρ, then the expected quality obtained by a customer with order ρ is $w[\rho] = w(N(\rho))$ if N is continuous at ρ, and otherwise

$$w[\rho] = \int_{N(\rho^-)}^{N(\rho)} w(r)\, dr/[N(\rho) - N(\rho^-)],$$

assuming random service within the class.

Allocations and Efficiency

An allocation specifies a subset of customers eligible for service, together with an assignment of these customers to priority classes. This assignment induces an assignment to service orders, randomized within each class. In turn, the associated distribution function N determines the quality associated with each priority class, using the above rules.

An efficient allocation maximizes the sum or integral of the customers' valuations of the resulting qualities. If the assignment produces a quality $w^\circ(v)$ for a customer whose type is $v \geq \underline{v}$ and no service for types $v < \underline{v}$, then this surplus is $\int_{\underline{v}}^1 u(v,w^\circ(v))dH(v)$. Using the assumptions above, if all priority classes are feasible then an efficient allocation serves all customers ($\underline{v} = 0$), and assigns customer type v the service order $r = \bar{H}(v)$. It is uniquely efficient if w is strictly decreasing. The case that only a limited number of priority classes is feasible is illustrated below.

Pricing

The role of pricing in priority service is to achieve an efficient allocation in the situation that each customer knows privately his type. That is, although we assume that the population's distribution of types is known, in practice it is necessary to induce customers to reveal their types indirectly via their selections of service conditions from a menu of options. There are several ways to construct such a menu, as discussed in Section 4. In one version, the options are described by pairs $\langle \rho, p[\rho] \rangle$, in which the price for the priority class ρ is $p[\rho]$; or if all service orders are feasible then an option can be specified as a pair $\langle r, p(r) \rangle$ indicating that the service order r has the price $p(r)$. In another version, an option is described by a pair $\langle w, P(w) \rangle$ specifying that $P(w)$ is the price of the quality w. In an optimal design, these prices induce customers to prefer their efficient service orders.

The formulation generalizes easily to the case that customers and the supplier subsequently observe the outcome of an additional random variable, say z, affecting preferences in the form that gross benefit for a type v customer is $u(v,w;z)$ and the quality obtained from service order r is $w(r;z)$. For the family of contingent options $\{\langle r,p(r;z)\rangle \mid z\}$ conditional on each outcome one can then charge the mean price $p(r) = E\{p(r;z)\}$. An example in the power context takes z to be the ambient temperature, which affects demand for heating and air conditioning. Similarly, one can aggregate over future dates. These trivial generalizations emphasize that the key elements are, first, the difference between the information known to customers (each knows his type) and the information known to the supplier

(the supply available), and second, the absence of a spot market to communicate and resolve these differences.

The construction of menus that achieve an efficient allocation is the topic of Section 3. We indicate occasionally the corresponding spot prices that could also achieve an efficient allocation were spot markets feasible. Note however that a spot market is essentially an algorithm to determine the efficient allocation in a particular contingency, whereas priority service elicits the efficient allocations for all contingencies simultaneously.

The Case of Several Priority Classes

To illustrate the design of priority service, we describe briefly the case that a finite number n of priority classes offered.

Suppose the menu consists of options $i = 1,\ldots,n$ described by pairs $\langle w_i, P_i \rangle$, where the service qualities w_i and the prices P_i increase with the index i. One index of quality is the expected quality: if q customers select an option, and Q select higher-numbered options, then

$$\bar{w}(q,Q) = \int_0^q w\,(Q+r)dr/q$$

is the expected quality, assuming random rationing among those customers selecting each option. Similarly, for a customer of type v, the gross benefit from the option can be specified as either $u(v,\bar{w}(q,Q))$ if only expected quality matters, or more generally as

$$\bar{u}(v;q,Q) = \int_0^q u(v,w(Q + r))\, dr/q.$$

If option i is selected by those types in an interval $[v_{i-1}, v_i]$ then their number is $q_i = H(v_i) - H(v_{i-1})$ and $Q_i = \bar{H}(v_i)$ is the number selecting higher priorities. In this case, consistency between the offered and expected qualities requires that $w_i = \bar{w}(q_i, Q_i)$. Further, $v_n = 1$ and for $i < n$ the type v_i at a boundary must be indifferent between adjacent options:

$$\bar{u}(v_i; q_i, Q_i) - P_i = \bar{u}(v_i; q_{i+1}, Q_{i+1}) - P_{i+1},$$

and $\bar{u}(v_0; q_1, Q_1) - P_1 = 0$ or $v_0 = 0$. If the types at the boundaries between classes are specified then these equations determine the qualities and all except one of the prices. Let U_i be the average gross benefit of customers within the i-th interval of types. Then an efficient design is obtained by choosing the boundary types v_i to maximize the total surplus $\sum_i U_i q_i$. We omit the conditions that characterize the solution, but some numerical examples are presented below.

The one degree of freedom in choosing the prices can be used in alternative ways. If it is mandatory that the least quality is served at the marginal cost of zero, then the constraint $P_1 = 0$ or $v_0 = 0$ is imposed; but if the choice of the lowest price P_1 is allowed to be optimized too, then typically it exceeds marginal cost and $v_0 > 0$. The reason is that serving the lowest-type customers degrades the quality available to higher-type customers in the lowest priority class; however, this consideration disappears as the number of classes increases.

• To illustrate, we use the standard example in which $u(v,w) = vw$, $H(v) = v$, $w(r) = 1 - r$, and $F(s) = s$. In the unconstrained case, the efficient boundary types are $v_i = [i + .5]/[n + .5]$. In the constrained case that the seller has an obligation to serve all customers, the efficient

boundary types are $v_i = i/n$. The constrained case yields the quantities, qualities, and prices $q_i = 1/n$, $w_i = [i - .5]/n$, and $p_i = P(w_i) = .5i[i-1]/n^2$. In both cases the total surplus increases to the limit 1/3 as the number n of classes increases, of which half represents the net revenue $\sum_i p_i q_i$ realized by the seller. The convergence rate is quadratic: with only four classes the total surplus for the two cases is already 0.329 and 0.328, respectively.

Similar results obtain in the case that the prices of priority classes are chosen to maximize the seller's net revenue. The essential difference is that the prices are all raised nearly uniformly, with a resulting contraction in the set of customers electing service. Thus, the main consequence of monopolist pricing is a reduction in the total surplus due to reduced market penetration.

• For the standard example cited above, and the cases with $n = 1,2$ and 4 priority classes, the middle column of Table 1 shows the optimal contingent prices $\hat{p}_i = p_i/w_i$ charged by a monopolist seller conditional on service delivery.[5] Allowed infinitely many classes, a monopolist uses a two-part tariff that charges a fixed subscription fee 1/8 in addition to the efficient priority prices, and thereby serves only those types exceeding $\underline{v} = 1/2$, resulting in the profit 5/24 and the total surplus 7/24.

Table 1
Priority Service by a Monopolist
Standard Example: $u(v,w)=vw$, $H(v)=v$, $w(r)=1 - r$

n	i	v_{i-1}	w_i	$\hat{p}_i$	q_i	$[U_i-p_i]q_i$	$p_i q_i$	$U_i q_i$
1	1	.577	.789	.577	.423	.070	.192	.263
2	2	.763	.882	.590	.237	.061	.123	.184
	1	.527	.645	.527	.237	.018	.080	.098
	Totals:				.473	.079	.204	.282
4	4	.877	.938	.604	.123	.039	.070	.109
	3	.754	.815	.563	.123	.025	.057	.082
	2	.630	.692	.529	.123	.014	.045	.059
	1	.507	.569	.507	.123	.004	.036	.040
	Totals:				.493	.082	.207	.289
∞	Totals:				.500	.083	.208	.292

The case that a continuum of qualities and service orders is offered corresponds to the limit as the number n of options increases to infinity. This is the case used in Section 3 to provide a brief exposition of the theory of priority service. However, we also show in Section 3.3 that the efficiency losses are small if a limited number of priority classes is offered.

We turn now to two examples that illustrate applications of priority service.

Example 1: Service Reliability and Interruption Costs

First we consider examples adapted to the capital-intensive peakload industries such as electric power. The missing spot market in such cases puts a price on access to service when supplies are scarce. Quality is interpreted as the reliability of service.

Interpret s as the available supply at some specified instant. Then,

the ratio r/s determines the customer's access to service when his service order is r. In particular, the technology yields the quality

$$w(r,s) = \begin{cases} 1 & \text{if } r/s \leq 1, \\ 0 & \text{if } r/s > 1, \end{cases}$$

when the service order is r and the supply is s. That is, the customer is served if supply is not exhausted when his service order is reached. If the type v is interpreted as the customer's value of service then his gross benefit is $u(v,w(r,s)) = vw(r,s)$. Recall that efficient rationing requires the service order $r = \bar{H}(v)$. The spot price $\pi(s)$ that implements this efficient allocation satisfies $s = \bar{H}(\pi(s))$: this spot price assures that only those s customers with the highest types obtain service. Alternatively, with random rationing every customer obtains the expected quality $\bar{w}(s) = \min\{s,1\}$.

Now suppose that supply is uncertain. Let $F(s)$ be the cumulative distribution function of supply. $\bar{F}(s) = 1 - F(s)$ indicates the probability that supply exceeds s. Random rationing yields the expected quality (i.e., reliability)

$$\bar{w} = \int_0^{1^-} s dF(s) + \bar{F}(1^-)$$

that is the mean effective supply per customer. Efficient rationing, on the other hand, yields a customer of type v the reliability $w^*(v) = w(\bar{H}(v))$, corresponding to the technology that specifies the quality as $w(r) = \bar{F}(r)$ to indicate that service is obtained only in the event that supply exceeds

the service order. There are several ways to achieve efficient rationing in this case. One is to charge the spot price $\pi(s)$ contingent on the supply that occurs. Another is to offer priority service in the form of contingent forward contracts $\langle r,p(r)\rangle$ that for a price $p(r)$ provide service if the supply exceeds the service order r. In Section 3.1 we demonstrate that a price system that induces customers to self-select the efficient service orders has the form:

$$p(r) = \int_r^1 \pi(s)dF(s).$$

That is, a customer is charged the expectation of the spot prices that would be paid for the same quality of service. The revenue collected by the seller is

$$\int_0^1 p(r)dr = \int_0^1 s\pi(s)dF(s).$$

which is precisely the expectation of the spot revenues.[6] Alternatively, if the customer pays a price $\hat{p}(r)$ only contingent on service then the expected payment is $p(r) = \bar{F}(r)\ \hat{p}(r)$ and this determines the contingent price $\hat{p}(r)$ as the conditional expectation of the spot price given that service is supplied.[7]

An alternative formulation interprets the customer's type v as the value of service per unit time. Quality is then interpreted as the expected duration of interrupted service. Consider a period in which more than one

event with deficient supply is unlikely; then the supply technology can be described by the distribution F of supply in the event of a shortfall and a function $t(s,s')$ that specifies the expected duration until an initially deficient supply s rises again above s'. The technology is summarized by the specification $w(r) = -\int_0^r t(s,r)dF(s)$ of the quality resulting from each service order r. With efficient rationing the customer again has the service order $r = \bar{H}(v)$ yielding the probability $F(\bar{H}(v))$ of interruption and a subsequent expected duration $t(s,\bar{H}(v))$ until resumption of service after an initially short supply $s < \bar{H}(v)$.

Example 2: Service Delays and Waiting Costs

Next consider examples adapted to make-to-order industries in which demand is backlogged in a queue for service. A spot market would allow customers to trade places in the queue of backlogged orders.

Interpreting the capacity or supply s as the service rate (customers served per day), the ratio r/s is the customer's service delay. If δ represents a common discount rate applied to the service delay, then one possible formulation of the technology represents the quality as

$$w(r,s) = e^{-\delta r/s},$$

interpreted as the discount factor used to compute the present value of service. With random rationing every customer obtains the expected quality

$$\bar{w}(s) = \int_0^1 w(r,s)dr = \frac{s}{\delta}\,[1-e^{-\delta/s}],$$

reflecting an equal chance of every service order. Efficient rationing requires that a customer with the valuation v obtains the priority $r = \bar{H}(v)$, so that customers with higher valuations are served earlier. This yields the quality $w(\bar{H}(v),s)$. A price system that charges $p(r)$ for the r-th service order implements this efficient allocation if

$$p'(r) = -w(r,s)\delta\bar{H}^{-1}(r)/s.$$

That is, the marginal saving from later service equals the marginal cost of the type efficiently served in that order. This differential equation determines the entire array of spot prices up to a constant of integration $p(0)$ that is the seller's price for immediate service.

In a similar example the type v represents the customer's cost of waiting. In this case the quality received is $w(r,s) = -r/s$.

Similar formulas apply if several competing firms offer service, with random rationing at each one. For example, if firm i has a service rate s_i and demand q_i at its price p_i, then the expected quality at the i-th firm is $\bar{w}_i = -\frac{1}{2} q_i/s_i$. Consequently, a customer with the waiting cost v prefers the firm for which the total cost $\frac{1}{2} vq_i/s_i + p_i$ is least, and it is this self-selection that determines the demands q_i at the firms. We examine competitive models of this kind in Section 5.

In the text we illustrate with a <u>standard example</u>. It is the special case of Example 1 (and one version of Example 2) in which $u(v,w) = vw$, $H(v) = v$, and $w(r) = 1 - r$. That is, customers' types and the supply are distributed uniformly and independently.

Related Work

The examples outlined above are the focus of recent work on priority service. Examples in which customers incur waiting costs from delayed service are studied by Reitman (1986). Mendelson and Whang (1986) study priority service to allocate service orders in an M/M/1 queue.

Examples in which customers forego service if interrupted are studied by Chao and Wilson (1987ab), Chao, Oren, Smith, and Wilson (1986ab, 1987), and Pitbladdo (1986). Of course, the large literature on peakload pricing considers such problems implicitly. Harris and Raviv (1981) derive a form of priority service offered by a monopolist seller having a fixed supply to serve a finite number of customers whose types are drawn independently from a probability distribution. They show that it is an optimal method of price discrimination based on differentiating the conditions of delivery. Applications to labor markets are developed by Mookerjee (1986).

3. BASIC THEORY OF PRIORITY SERVICE

This section sketches the main theoretical aspects of priority service. We emphasize the price system's role to induce customers to select efficient service orders. The case of competing firms who set their prices to maximize expected profits is deferred to Section 5.

Recall that each customer's type is described by a number v, and $\bar{H}(v)$ is the number of customers whose types are not less than v. A customer of type v obtains the gross benefit $u(v,w)$ from the service quality w, where u is increasing in both arguments and has a cross-partial $u_{vw} > 0$. The technology specifies that the quality of service is a decreasing

function $w(r)$ of the service order r assigned to the customer. Allocative efficiency therefore requires that type v receives the service order $r = \bar{H}(v)$. A schedule $\langle r,p(r)\rangle$ of service options specifies a price $p(r)$ for each service order r. Consequently, a customer whose type is v prefers the option for which the net benefit $u(v,w(r)) - p(r)$ is maximal. Alternatively, options can be specified as pairs $\langle w,P(w)\rangle$ of qualities and corresponding prices.

3.1 PRICING PRIORITY SERVICE

We first describe how service options are priced. The prices must be chosen to induce efficient rationing, anticipating that each customer selects a preferred option based on his privately known type. We derive these prices for an arbitrary menu and then we show how the results are adapted to construct an optimal menu that implements an efficient allocation.

Given a menu W of options $\langle w,P(w)\rangle$ specifying qualities and their prices, the net benefit obtainable by a customer of type v is

$$U(v) = \max\{0,\max_{W}\{u(v,w) - P(w)\}\}.$$

Let $w^{o}(v)$ indicate the quality chosen by a type $v \geq \underline{v}$ who selects some option from the menu, where $\underline{v}$ is the least type electing service. We show that customers' freedom to select qualities according to the price schedule implies that customers getting higher quality must pay an amount sufficient to compensate all those customers with lesser types for the resulting degradation of the qualities that they receive.

PROPOSITION 1. Customers' selection of service qualities implies that the price system satisfies:

$$P(w^\circ(v)) = P(w^\circ(\underline{v})) + \int_{\underline{v}}^{v} u_w(x,w^\circ(x))dw^\circ(x).$$

PROOF. Our assumptions imply that w° is necessarily a nondecreasing function of the type v. The standard methods used in the theory of self-selection therefore verify the validity of the envelope theorem: $U'(v) = u_v(v,w^\circ(v))$ for almost all types $v \geq \underline{v}$. Integrating this differential equation yields

$$U(v) = U(\underline{v}) + \int_{\underline{v}}^{v} u_v(x,w^\circ(x))dx,$$

where either $\underline{v} = 0$ or $U(\underline{v}) = 0$. Integrating by parts and using the definition of $U(v)$, one obtains the formula stated in the Proposition. Q.E.D.

We now apply this general result to the design of an optimal menu. Efficient rationing requires that type v obtains the service order $r = \bar{H}(v)$ and thereby the quality $w^\circ(v) = w(\bar{H}(v))$; consequently, changing the variable to the service order $r = \bar{H}(v)$ provides a formula for the efficient prices in terms of the service order:

$$p(r) = P(w(r)) = p(\underline{r}) + \int_{\underline{r}}^{r} u_w\ (\bar{H}^{-1}(t),\ w(t))dw(t),$$

where $\bar{r} = \bar{H}(\underline{v})$ is the last order served. A further requirement for efficiency is that all types $v \geq 0$ are served. Consequently, $\underline{v} = 0$, $\bar{r} = 1$, and $p(\bar{r}) = 0$ suffices. If this price system is offered, then all customers prefer to elect service and they have incentives to choose the efficient service orders. It is also the price system that results from offering the service orders in an auction: each increment in priority commands an incremental price from the winning bidder that is the marginal benefit of the resulting increment in quality to other customers who win adjacent priorities.

We remark that if the quality technology is modeled as the bivariate function $w(r,s)$ depending on a stochastic supply s, then the analogous formula applies:

$$p(r) = p(\bar{r}) + \int_0^{\infty} \int_{\bar{r}}^{r} u_w(\bar{H}^{-1}(T), w(t,s))\, dw(t,s)\, dF(s),$$

but we do not pursue this level of generality.

- In the context of Example 1 where $w(r) = \bar{F}(r)$:

$$p(r) = P(\bar{F}(r)) = \int_r^1 \bar{H}^{-1}(s) dF(s).$$

This is the formula displayed in Example 1 of Section 2 where we noted its interpretation in terms of the expectation of the spot price $\pi(s) = \bar{H}^{-1}(s)$ for comparable quality. In the standard example, $p(r) = p(\bar{r}) + \frac{1}{2}[1 - r]^2$,

and $p(\bar{r}) = 0$ is further required for efficiency.

The construction is similar if the technology specifies both the probability distribution $F(s)$ of an initially short supply s and the expected time $t(s,s')$ to bring the supply up to a level s'. Assume that $t(s,s) = 0$ and that t decreases in s and increases in s'. In this case, the quality provided by a service order r is expressed in terms of the expected duration $w(r) = -\int_0^r t(s,r)dF(s)$ of interrupted service. The optimal price schedule is again

$$p(r) = \int_r^1 \bar{H}^{-1}(x)\, dw(x),$$

which can be interpreted in terms of the expected spot price for comparable quality by changing the variable of integration to the spot price π via $x = \bar{H}(\pi)$.

Monopolistic Pricing and the Seller's Revenue

The implications of Proposition 1 extend directly to the case of a profit-maximizing monopolist. If a monopolist serves types $v \geq \underline{v}$, then its profit is

$$\int_{\underline{v}}^1 P(w^\circ(v))\, dH(v) = u(\underline{v},\, w(\bar{r}))\bar{r} - \int_0^{\bar{r}} r u_w\, (\bar{H}^{-1}(r), w\,(r))dr,$$

where the lowest type $\underline{v}$ receives service order $\bar{r} = \bar{H}(\underline{v})$. The condition $p(\bar{r}) = u(\underline{v},\, w(\bar{r}))$ determines this lowest type. The optimal price that is charged for this last service order is identified by the further condition that

$$\underline{v} = \arg\max_x \{u(x, w(\bar{r})\bar{H}(x)\}.$$

The monopolist restricts demand by setting the base price $p(\bar{r})$ above marginal cost.

- If $u(v,w) = vw$ then $\underline{v}$ is chosen to maximize $\underline{v}\,\bar{H}(\underline{v})$. In the standard example this yields $\underline{v} = 1/2$ and $\bar{r} = 1/2$. The resulting price schedule is $p(r) = 1/8 + \frac{1}{2}[1-r]^2$, which yields the profit 5/24. In contrast, the efficient price schedule that serves all customers is $p(r) = \frac{1}{2}[1 - r]^2$, which yields the smaller profit 4/24.

In practice, state enterprises must collect sufficient revenue to recover their costs of capital. In the United States, public utilities impose, in addition to an energy charge reflecting marginal cost, a 'demand charge' for this purpose. With priority service, all or part of the demand charge can be replaced by priority service charges. Additional charges might be imposed, nevertheless, to recover sunk capital costs and to cover administrative costs. If the seller's revenue constraint is binding then additional terms enter the formulation via an augmented Lagrangian expression. The net effect of this alteration is to induce the seller to exploit its monopoly power sufficiently to raise the required revenue. As seen above, a monopolist adheres to the efficient service order, but serves fewer customers by imposing an additional fixed subscription fee. The same applies to a public utility constrained by a revenue requirement.

Multiple Priorities

The simplicity of the pricing formula is due partly to the assumption that customers' types are described by a one-dimensional parameter.

Multiple dimensions are important in some context because the technology of supply involves a cascade of queues for service in each of which customers are subject to rationing. In the case of a power interruption with delayed resumption, for example, a customer first enters a queue for uninterrupted service, and failing service there, then enters a queue for resumed service as the supply subsequently increases. That is, he obtains service at the first queue in which his assigned priority is sufficiently high. Unfortunately, the simple formulas for the case of one-dimensional types do not always carry over neatly to more general cases with these complicating features. To illustrate, we describe an example that succinctly summarizes the difficulty,

Suppose that a customer's type is a pair (c,v) indicating that he incurs a fixed shutdown and restart cost c if interrupted, and a further foregone service value v for each minute until resumption. Service quality is represented by a pair $\langle w_1, w_2 \rangle$ specifying the chance w_1 that service is not interrupted, and following an interruption, the expected duration w_2 until service is resumed. Efficient rationing in this case requires two service orders for each customer, say $\langle r_1, r_2 \rangle$. Service is interrupted if the initially short supply s is less than r_1 and then resumed when the supply rises above r_2. With this technology, these service orders provide the qualities $w_1(r_1) = \bar{F}(r_1)$ and

$$w_2(r_2; r_1) = -\int_0^{r_1} t(s, r_2 + \Delta(s,r_2))\, dF(s),$$

where $\Delta(s,r)$ is the measure of customers assigned interruption orders between s and r and resumption orders exceeding r — namely, those who obtain service after an interruption only because they were not interrupted

initially. For a customer of type (c,v), efficient rationing requires that the service order for resumption is $r_2 = \bar{H}(v)$; the service order for interruption is obtained by solving a complicated optimization problem. To implement these efficient service orders via customers' self-selection, it must be that in choosing among the menu of options $\langle r_1, r_2; p(r_1,r_2)\rangle$ to obtain the greatest expected benefit $cw_1(r_1) + vw_2(r_1, r_2) - p(r_1, r_2)$, the customer of type (c,v) prefers his efficient service orders.

Oren (1987) shows that there need not exist a price schedule that meets this requirement. Technically, the difficulty arises because the differential equations that characterize a price schedule need not satisfy an integrability condition. The result is plausible from the fact that at a time after an interruption the imputed spot price depends on the current supply s' and also on the initial supply s, via the measure $\Delta(s,s')$ of uninterrupted customers whose valuations are less than the current spot price. That is, the spot-price process depends on history via the initial event that precipitates rationing of supplies. It is clear that in such cases priority service charges can not be calculated simply in terms of the expected spot prices for comparable qualities. Thus, we caution that the formula for priority pricing need not generalize (and indeed need not exist) for more complicated models in which self-selection precludes full efficiency.

3.2 WELFARE PROPERTIES OF PRIORITY SERVICE

The principal feature of priority service is that it uses contingent forward contracts conditioned on service orders to supplant spot markets.

The efficiency gains from spot markets are well known, and priority service realizes essentially these same gains if a sufficiently rich menu is offered.

The distributional consequences of priority service are also important. The following result states that a simple <u>equal</u> redistribution of the incremental revenues raised by shifting from randomly ordered service to efficient priority service is sufficient to assure that no customer presently served is disadvantaged by the change. We assume that a continuum of service orders is offered as the menu, but comparable results obtain if the number of priority classes is limited.

PROPOSITION 2. Priority service is Pareto superior to randomly ordered service if incremental revenues are redistributed equally to present customers.

PROOF. If randomly ordered service is offered at a nonnegative price $\bar{p}$ then those types $v \geq \underline{v}$ subscribe, where $\bar{p} = \int_0^{\bar{r}} u(\underline{v}, w(r))\, dr/\bar{r}$ and $\bar{r} = \bar{H}(\underline{v})$. The seller's revenue is $\overline{pr}$, and a customer of type $v \geq \underline{v}$ receives the net benefit

$$\begin{aligned}
\bar{U}(v) &= \int_0^{\bar{r}} u(v,w(r))\, dr/\bar{r} - \bar{p}, \\
&= \int_1^{\underline{v}} u(v,w(\bar{H}(x)))\, d\bar{H}(x)/\bar{H}(\underline{v}) - \bar{p}, \\
&= u(v,w^{\circ}(\underline{v})) + \int_{\underline{v}}^1 \bar{H}(x)\, u_w(v,w^{\circ}(x))dw^{\circ}(x)/\bar{H}(\underline{v}) - \bar{p},
\end{aligned}$$

where $w^{\circ}(v) = w(\bar{H}(v))$ denotes the efficient quality assignment for type v. On the other hand, with efficient priority service, if type v subscribes

then his service order is $r = \bar{H}(v)$, the price he pays is

$$P(w^\circ(v)) = P(w^\circ(\underline{v})) + \int_{\underline{v}}^{v} u_w(x,w^\circ(x))dw^\circ(x),$$

and his net benefit is

$$\begin{aligned} U(v) &= u(v,w^\circ(v)) - P(w^\circ(v)), \\ &= u(v,w^\circ(v)) - P(w^\circ(\underline{v})) - \int_{\underline{v}}^{v} u_w(x,w^\circ(x))\, dw^\circ(x). \end{aligned}$$

With incremental revenues redistributed equally to those types $v \geq \underline{v}$ who subscribe to randomly ordered service, in amounts such that the seller's net revenue is still $\overline{pr}$, the net price to type $\underline{v}$ satisfies:

$$\begin{aligned} P(w^\circ(\underline{v})) &= -\int_{\underline{v}}^{1} \int_{\underline{v}}^{v} u_w(x,w^\circ(x))dw^\circ(x)\; dH(v)/\bar{H}(\underline{v}) + \overline{pr}/\bar{H}(\underline{v}), \\ &= -\int_{\underline{v}}^{1} L(x)u_w(x,w^\circ(x))dw^\circ(x) + \bar{p}, \end{aligned}$$

where $L(x) = \bar{H}(x)/\bar{H}(\underline{v})$. This equality holds provided that all types $v \geq \underline{v}$ subscribe to priority service, which we verify below. For those types $v \geq \underline{v}$, the gain $G(v) = U(v) - \bar{U}(v)$ satisfies:

$$
\begin{aligned}
G(v) &= \{u(v,w^\circ(v)) + \int_{\underline{v}}^{1} L(x)u_w(x,w^\circ(x))dw^\circ(x) - \int_{\underline{v}}^{u} u_w(x,w^\circ(x))dw^\circ(x)\} \\
&\quad - \{u(v,w^\circ(\underline{v})) + \int_{\underline{v}}^{1} L(x)u_w(v,w^\circ(x))dw^\circ(x)\}, \\
&= u(v,w^\circ(v)) - u(v,w^\circ(\underline{v})) - \int_{\underline{v}}^{v} u_w(x,w^\circ(x))dw^\circ(x) \\
&\quad + \int_{\underline{v}}^{1} L(x)[u_w(x,w^\circ(x)) - u_w(v,w^\circ(x))]dw^\circ(x), \\
&\geq u(v,w^\circ(v)) - u(v,w^\circ(\underline{v})) \\
&\quad - \int_{\underline{v}}^{v} [L(x)u_w(v,w^\circ(x)) + (1 - L(x))u_w(x,w^\circ(x))]dw^\circ(x), \\
&\geq u(v,w^\circ(v)) - u(v,w^\circ(\underline{v})) - \int_{\underline{v}}^{v} u_w(v,w^\circ(x))dw^\circ(x), \\
&= 0.
\end{aligned}
$$

In these relations, the first inequality uses the cross-partial assumption that if $x > v$ then $u_w(x,w) \geq u_w(v,w)$; and the second uses the property of a convex combination of two terms that it is less than the larger term. Thus, we conclude that all types $v \geq \underline{v}$ gain from the adoption of priority service - and therefore they all subscribe, since they subscribe to randomly ordered service. Lower types (who refuse randomly ordered service) may also gain by electing priority service, and the seller may have increased revenues, if there are types $v < \underline{v}$ for which $u(v, w^\circ(v)) > P(w^\circ(v)) > 0$ - assuming that qualities $w^\circ(v)$ for which $v > \underline{v}$ and $P(w^\circ(v)) < 0$ are excluded from service. Thus, this implementation of priority service is Pareto superior to randomly ordered service. Q.E.D.

Proposition 2 implies that adoption of priority service can benefit every customer without reducing the seller's revenue. It suffices to refund

only a portion of the revenues sufficient to prevent any customer from being disadvantaged and then using the remainder to reduce assessments for capital recovery or for capacity expansion to improve qualities.

• To illustrate, in the standard example assume that $\bar{p} = 0$ so that $\underline{v} = 0$ and $\bar{r} = 1$. Then randomly ordered service yields the customer of type v the net benefit v/2, whereas efficient priority service yields the net benefit $v^2/2 - p(1)$. Taking $p(1) = -1/6$ so that net revenues of the seller are zero, the customer's gain from priority service is $G(v) = [v^2 - v + 1/3]/2$. Type $v = 1/2$ benefits the least, but still the gain $G(1/2) = 1/24$ is positive. The average gain per customer is 2/24.

The distributional effects of introducing priority service are similar if customers adapt their end-use technologies to the availability of priority service, but the gains are larger. Elaborating the previous example, suppose that a customer of type v can obtain a service value t at the cost of an investment $t^2/4v$, so that the customer's gross benefit is $u(v,w) = tw - t^2/4v$ if he obtains quality w. With randomly ordered service the only available quality is $w = 1/2$; therefore, the customer chooses the technology $t(v) = v$ and obtains the net benefit v/4, including investment costs. With this choice of technology fixed, priority service yields the same gains stated above. But, if the customer optimizes his technology choice, namely $t(v) = 2vw^\circ(v)$ using the efficient quality selection $w^\circ(v) = v$ induced by priority service, then the gain for a customer of type v is $G(v) = v^3/3 - v/4 + 1/6$. The least gain in this case is $G(1/2) = 2/24$. The average gain per customer is 3/24, which is 50% larger than previously. This reflects the general principle that an added advantage of product differentiation is the opportunity allowed customers to adapt their end-use

technologies to their diverse quality selections. In the present example, this opportunity is manifest in the induced gross benefit function $u(v,w) = vw^2$ that results from optimizing the choice of technology for each quality w. These additional gains are not discernible in demand data obtained in the absence of priority service.

3.3 MENU VARIETY

A striking feature of numerical examples is that a few priority classes suffice to obtain most of the gains from priority service. This complements the observation that in practice differentiation of priority services is rather coarse. The generality of this feature is established next. The proof assumes that u, w, and $\bar{H}^{-1}$ have uniformly bounded second derivatives.

PROPOSITION 3. Priority service with n priority classes incurs an efficiency loss that is of order $1/n^2$.

PROOF. It suffices to prove the stronger statement that if the priority classes have <u>equal</u> numbers of customers then the efficiency loss is $0(1/n^2)$. Let $S_n = \sum_i S_n^i$ be the total surplus when there are n equal-sized classes $i = 1,\ldots,n$ comprising customers of types $v \in [v_{i-1}, v_i]$, where $v_0 = 0$ and $v_n = 1$ Let $Q_i = \bar{H}(v_i)$ so that $\Delta = Q_{i-1} - Q_i$ is the fraction of customers in each class. Also let $u_i = u(v_i, w(\bar{H}(v_i)))$ and, defining $v(x) = \bar{H}^{-1}(x)$, let $u_i' = du(v(x),w(x))/dx$ evaluated at $x = Q_i$. The surplus realized by the i-th priority class is

$$S_n^i = \int_{Q_i}^{Q_i+\Delta} \int_0^{\Delta} u(v(x),w(Q_i + r))dr\, dx/\Delta$$

$$= \Delta u_i + \frac{1}{2}\Delta^2 u_i' + 0(\Delta^3),$$

where the second equality results from a Taylor series expansion of the numerator around $\Delta = 0$. The potential surplus with infinitely many classes is

$$S_\infty = \int_0^1 u(v,w(\bar{H}(v)))\ dH(v) = \Delta \sum_i u_i + \frac{1}{2}\Delta[u_0 - u_n] + 0\ (\Delta^2),$$

where the second equality states the trapezoid rule for numerical integration. Combining these expressions, the efficiency loss is

$$\mathcal{L}_n = S_\infty - S_n = \frac{1}{2}\Delta\ \{u_0 - u_n - \Delta \sum_i u_i'\} + O(\Delta^2) = 0\ (\Delta^2),$$

where the last equality follows from the observation that the quantity in curly brackets is $O(\Delta)$, according to the trapezoid rule applied to the integral

$$\int_0^1 \frac{d}{dx}\, u(v(x),w(x)\ dx = u_0 - u_n.$$

Because $\Delta = 1/n$, this proves that the efficiency loss is of order $1/n^2$.

Q.E.D.

A corollary of this proof is that it is asymptotically efficient to use priority classes of equal sizes. For the special case that $u(v,w) = vw$, Proposition 3 can be strengthened to obtain a convenient bound on the rate of convergence. When n priority classes $i = 0,\ldots,$ n-1 are offered the attainable surplus can be written as $S_n^\circ = \sum_{i=0}^{n-1} \bar{v}_i \bar{w}_i \Delta_i$, assuming that supplies are rationed randomly within the lowest-priority class served in each event, which is class i if the supply x is in the interval (x_i, x_{i+1}). In this formula, $x_0 = 0$, $x_n = 1$, $\Delta_i = x_{i+1} - x_i$, and within the i-th such interval the average valuation and average service quality are

$$\bar{v}_i = \int_{x_i}^{x_i+\Delta_i} v(x)\, dx/\Delta_i \quad \text{and} \quad \bar{w}_i = \int_{x_i}^{x_i+\Delta_i} w(x)\, dx/\Delta_i .$$

The average within-class variances are $\sigma_n^2 = \sum_i s_i^2\Delta_i$, and $\rho_n^2 = \sum_i r_i^2\Delta_i$ where

$$s_i^2 = \int_{x_i}^{x_i+\Delta_i} [v(x)-\bar{v}_i]^2\, dx/\Delta_i \quad \text{and} \quad r_i^2 = \int_{x_i}^{x_i+\Delta_i} [w(x) - \bar{w}_i]^2\, dx/\Delta_i .$$

COROLLARY. If there are n priority classes then $\mathcal{L}_n \le \sigma_n\rho_n$. Moreover, if the class sizes are equal then σ_n and ρ_n are each of order $1/n$.

PROOF. Using the Schwartz inequality:

$$\int_{x_i}^{x_i+\Delta_i} v(x)w(x)dx/\Delta_i - \bar{v}_i\bar{w}_i = \int_{x_i}^{x_i+\Delta_i} [v(x)-\bar{v}_i][w(x)-\bar{w}_i]dx/\Delta_i$$

$$\le \left(\int_{x_i}^{x_i+\Delta_i} [v(x)-\bar{v}_i]^2 dx/\Delta_i \right)^{1/2} \left(\int_{x_i}^{x_i+\Delta_i} [w(x)-\bar{w}_i]^2 dx/\Delta_i \right)^{1/2} = s_i r_i .$$

Consequently,

$$S_\infty = \sum_i \int_{x_i}^{x_i+\Delta_i} v(x)w(x)dx \le \sum_i [s_i r_i + \bar{v}_i\bar{w}_i]\Delta_i = \sum_i s_i r_i \Delta_i + S_n^\circ ,$$

and $$\mathcal{L}_n \le \sum_i s_i r_i \Delta_i \le \sqrt{\left(\sum_i s_i^2\Delta_i\right)\left(\sum_i r_i^2\Delta_i\right)} = \sigma_n\rho_n ,$$

using the Cauchy inequality. This is the claimed bound on the efficiency loss. The next task is to verify that $n\sigma_n$ and $n\rho_n$ are bounded as n increases, repeating the method in the proof of the Proposition. It suffices to prove the result only for σ_n. Using a Taylor series approximation for $\bar{v}_i$,

$$\bar{v}_i^2 = v(x_i)^2 + v(x_i)v'(x_i)\Delta_i + O(\Delta_i^2).$$

Consequently, assuming equal classes $\Delta_i = \Delta$,

$$\Delta \sum_i \overline{v}_i^2 = [\Delta \sum_i v(x_i)^2] + \frac{1}{2}\Delta[\Delta \sum_i 2v(x_i)v'(x_i)] + O(\Delta^2)$$

$$= \left[\int_0^1 v(x)^2 \, dx - \frac{1}{2} \Delta [v(1)^2 - v(0)^2] + O(\Delta^2) \right]$$

$$+ \frac{1}{2} \Delta [v(1)^2 - v(0)^2 + O(\Delta)] + O(\Delta^2)$$

$$= \int_0^1 v(x)^2 \, dx + O(\Delta^2),$$

where the second quality uses the trapezoid rule in each bracketed term. Thus we have shown that

$$\sigma_n^2 = \int_0^1 v(x)^2 \, dx - \sum_i \overline{v}_i^2 \Delta_i = O(\Delta^2),$$

which is the claimed result given that $\Delta = 1/n$. Q.E.D.

This bound is tight, since $\mathcal{L}_n = \sigma_n \rho_n = 1/12n^2$ for the standard example. In the special case that $n = 1$, $\sigma_1 \rho_1$ is an upper bound on the gain $\mathcal{L}_1$ from introducing priority service, as compared to randomly ordered service within a single class. In contexts such as Example 1, it is useful to decompose ρ_1^2 into the sum $\rho_1^2 = \sigma_q^2 + \tau$ of the variance σ_q^2 of the effective supply $q = \min\{1,s\}$ and a parameter τ peculiar to the supply distribution:

$$\sigma_q^2 = 2 \int_0^1 sw(s)ds - \left[\int_0^1 w(s)ds \right]^2,$$

$$\tau = \int_0^1 [w(s) - s]^2 \, ds - 1/3.$$

For the standard example, $\tau = 0$, and generally $\tau \leq w(0)^3/4$.

These arguments do not apply to a profit-maximizing firm because such a

firm does not choose the priority prices to maximize total surplus. We omit the proof that in this case the profit loss from using n classes is generally of order 1/n. Table 1 and other numerical examples indicate, nevertheless, that in some cases the convergence is still quite rapid.

3.4 PRIORITY INSURANCE

In applications like Example 1, priority service imposes risks of service interruptions on each customer; indeed, riskiness of supply is a main motivation for adopting priority service. If customers are risk averse, then full efficiency requires that risks are shared efficiently among the customers and the firm. In important applications such as power, a state enterprise or public utility is much less risk averse than each customer. Consequently, we investigate the case that the firm or a private underwriter offers compensatory insurance against the risk of loss from service interruptions, and does so at actuarially fair rates.

The principles involved are simple, so it suffices to address a variant of Example 1. Describe a customer's type by a pair (v,V) in which v is the value of service and V is the customer's von Neumann-Morgenstern utility function defined on net benefits. Interpret -w(r,s) as the duration of an interruption parameterized by an initial deficient supply s. In the menu include supplementary insurance options of the form $\langle y_r, \tilde{p}[y_r] \rangle$ in which y_r is a function of the supply that specifies the compensation paid in the event of interruption for service order r:

$$y_r(s) = \begin{cases} |w(r,s)| & \text{if } s < r, \\ 0 & \text{if } s \geq r. \end{cases}$$

If the customer of type (v,V) purchases the priority service option

$\langle r,p(r)\rangle$ and supplements it with x units of the insurance option $\langle y_r,\tilde{p}[y_r]\rangle$ then his total charge is $P(r,x) = p(r) + x\tilde{p}[y_r]$ and his expected utility is

$$\bar{V}(r,x) = \int_0^r V([x-v]|w(r,s)| - P(r,x))\, dF(s) + V(-P(r,x))\bar{F}(r).$$

Assume that the insurance premium is the actuarial value of the contingent compensation:

$$\tilde{p}[y_r] = \int_0^\infty y_r(s)dF(s) = \int_0^r |w(r,s)|dF(s).$$

Assume also that the priority service charge is the efficient one calculated previously in the absence of risk aversion: it satisfies

$$p'(r) = -\bar{H}^{-1}(r)\, d\tilde{p}[y_r]/dr,$$

and therefore,

$$P(r,\bar{H}^{-1}(r)) = p(1) + \int_1^r \tilde{p}[y_\rho]d\bar{H}^{-1}(\rho).$$

It is straightforward to verify that the choices that maximize this expected utility are the efficient ones. The customer selects the efficient service order $r = \bar{H}(v)$, and he obtains full coverage $x = v$ of his risk.

• To illustrate, in the standard example $p(r) = p(1) + \frac{1}{2}[1-r]^2$ and $\tilde{p}[y_r] = F(r) = r$; therefore, the total charge $P(r,\bar{H}^{-1}(r)) = p(1) + \frac{1}{2}[1-r^2]$ is made for the coverage that provides compensation $1 - r$ in the event $s < r$ that the customer is interrupted, anticipating that a customer of type $v=1-r$

will select this coverage. Note that $dP(r,\bar{H}^{-1}(r))/dr = -r$, indicating that each increment in the service order incurs an increment in the total charge that is the actuarial value of the incremental coverage.

This construction generalizes readily. A customer selecting the priority class r is offered supplementary insurance that provides compensation $C-u(\bar{H}^{-1}(r),w(r,s))$ in the event s, where C is an arbitrary constant. If the premium is actuarially fair then the customer prefers the full coverage $x = 1$. Or, if the premium and the priority charge are bundled together, then the customer still prefers to select the efficient service order.

This result illustrates two general principles. First, in response to actuarially fair insurance rates, each customer prefers full coverage of his risk. And, second, given that full insurance coverage equates marginal utility in all contingencies, each customer prefers to minimize his total payment: if the priority service charge is calculated as in the absence of risk aversion, this results in the efficient service order. The total payment assesses the actuarial value of the incremental insurance coverage for each increment in the service order; that is, no cross subsidization among risk classes is required.

When a customer's risk is fully insured, he is indifferent whether he is interrupted. The firm or underwriter now bears all the financial risk, and it is useful to examine their incentives to follow the efficient service order. Given that some number of customers must be interrupted, a privately-owned firm prefers to interrupt those customers to whom it must pay the smallest compensations. In fact, this rule yields the efficient service order. Thus, it suffices that the firm offers all varieties of

compensatory insurance, and then chooses the service order to minimize the compensation it pays out. If the rates for incremental coverage are actuarially fair, involving no cross subsidization among risk classes, then an efficient allocation results.

3.5 CAPACITY PLANNING

A persistent difficulty in the management of state enterprises is determination of optimal capacity. The problem stems from inadequate information. For example, in the United States, public utilities lack market data that reveal customers' aggregate willingness to pay for investments in capacity. We mention briefly how priority service alleviates the problem.

The source of the difficulty is that the quality attributes of capacity are unpriced in the market. In the case of power, capacity increments improve mainly the reliability of power supply. Yet, if only a single class of service is offered then customers' demands reveal little about their valuations of reliability. Differentiation of service reliability into several priority classes, on the other hand, provides direct evidence about customers' willingness to pay for the quality improvements that would derive from capacity expansion.

To illustrate, suppose that a continuum of service orders is offered, and assume that the quality (eg., reliability) associated with service order r is represented as $w(r;k)$, indicating its dependence on the installed capacity k. Assume that the quality w is a differentiable increasing function of the capacity k. If scarce supplies are rationed efficiently, and the unit cost of capacity is c, then the aggregate benefit net of

capacity costs is

$$\int_0^1 u(v(r),w(r;k))\,dr - ck,$$

where $v(r) = \bar{H}^{-1}(r)$ is the type of the customer who obtains service order r. The necessary condition for the optimality of marginally increasing capacity beyond the level k is therefore that

$$\int_0^1 u_w(v(r),w(r;k))w_k(r;k)dr > c.$$

The information about customers required to evaluate this criterion is provided by customers' market behavior in response to priority service. Recall that

$$p'(r) = u_w(v(r),w(r;k))w_r(r;k),$$

which enables the seller to infer the marginal valuation of quality from the ratio of the marginal priority charge and the marginal quality from a lower service order.

For instance, suppose that $w_k(r;k) = -w_r(r;k)$ if it is the difference k - r that matters; i.e., a capacity increment advances the effective service orders of all customers. In this special case the criterion for expansion of capacity is merely that $p(0) - p(1) > c$. That is, capacity is expanded if the priority charge of the highest-priority customer is sufficient to pay for it. This follows directly from the property of the priority pricing schedule that each customer pays a charge sufficient to compensate lower-priority customers for the resulting degradation of the qualities of their services; consequently, the maximum charge also measures the aggregate willingness to pay for quality improvement.

A spot market could also provide such data if spot prices could be observed in sufficiently many contingencies in a stationary environment to estimate their distribution. A striking feature of priority service is that it elicits the entire distribution of spot prices ex ante. For example, the 'priority points' scheme described below in Section 4 provides this distribution without any informational burden on the seller.

4. IMPLEMENTATION OF PRIORITY SERVICE

In this section we explore the implementation of priority service in practical situations. A state enterprise or public utility is assumed. We concentrate on schemes that reduce the informational requirements imposed on the seller. The context used is Example 1 in which quality is synonymous with reliability of service, as in the case of electric power.

Menu Parameterization

Within the formulation of the model, alternative descriptions of the options in the menu are equivalent. In practice, however, they vary significantly in terms of the information required by the seller and by a customer. In turn, these requirements affect the design of contracts and the organization of the market. Some possible parameterizations of options are the following.

[1] Quality w. Each customer compares quality w and price P(w). Alternatively a customer could submit his type v and be charged $P(w^*(v))$ for the promised quality $w^*(v)$. The seller adjusts prices to achieve the promised qualities. This scheme imposes all the informational requirements on the seller.

[2] Service order or priority class r. The customer compares quality w(r) and price p(r). The quality assessment requires knowledge of the supply distribution and prediction of the number of customers selecting higher priority classes. The seller adjusts prices to obtain efficient self-selection by customers, which requires estimation of the

distribution of customers' types. Alternatively, the seller estimates the distribution of spot prices $\pi(s)$. In other variations the customer is billed ex post for the spot price $\pi(s)$ after service is provided in the contingency $r \leq s$, or if he chooses a reservation price v then in the contingency $v \geq \pi(s)$.

[3] Price p or contingent price $\hat{p}$. The seller merely serves customer in order of the prices selected. A customer must assess the qualities expected from alternative bids. This scheme imposes the entire informational burden on customers.

This short list of possibilities demonstrates that the loci of informational requirements can be placed anywhere in the spectrum between the seller and the customer. Admittedly information about the distribution of customers' types is revealed by their selections, but this does not aid in an initial implementation, which can be risky for the seller (in version [1]) or customers (in [3]). And in any case there are continuing difficulties tracking changes in the distribution, due for example to altered investments in end-use appliances as mentioned in Section 3.2. The choice among the various schemes depends primarily on whether the informational burden is economically assigned to the seller or the customers.

Contracting and enforcement are also diverse; e.g., in [1] promised qualities (reliabilities) are difficult for customers to verify. In [2] the seller replicates indirectly the operation of the absent spot market. With uncertain demand, however, the financial consequences can differ according to whether the seller's expected revenue is the product of the expected spot price and the expected quantity sold, or the expectation of the product of the spot price and the quantity sold. In [3] the seller is absolved of all responsibility other than to observe the rationing order.

Equally diverse are the possibilities for organizing the market for priorities. We mentioned in Section 3.4 the extreme form of [1] in which

the seller, via priority insurance, assumes both the risks and the burden of information collection. At the other extreme, in one version of [3] the seller offers for sale an unlimited supply of 'priority points' at a price of $1 each: customers are then served in reverse order of the number of points purchased. Alternatively, if the supply of priority points is fixed then their price equilibrates demand and supply, possibly via an auction or a brokerage market that allows a role for intermediaries who specialize in acquiring the requisite information.

Partial Implementation

Priority service can be implemented partially without major modifications. We mention two examples.

First suppose that priority service is not offered to some market segment, such as residential customers. This segment then effectively becomes a special priority class comprising undifferentiated customers, all of whose service valuations are taken to be the average of the valuations within the class, and this average valuation determines its efficient service order. If this service order lies within the range of service orders selected by other customers then low-priority customers incur interruptions before this special class, and high-priority customers, only after this special class has been entirely interrupted.

Second, suppose that some customers interrupt their service automatically based on remote sensors, in the case of customers who agree to curtail their load whenever the ambient temperature exceeds a threshold, or in an alternative current system, whenever the line frequency drops below a designated threshold level. A similar case includes air conditioners,

heaters, and pumps that 'cycle' between on and off to reduce the load. In these cases it suffices to compute the conditional expectation of the spot price given that the sensor exceeds the threshold, and then to use this and the distribution of the sensor's values to compute the priority service charge in terms of the expected spot price. In general, this characterization of the priority service charge can be used to establish the terms for a variety of interruptible service contracts.

Optimal Contract Period

Absent imperfect communication and costly transactions, a spot market is efficient. Otherwise, forward contingent contracts can economize on monitoring, communication, and transaction costs. The chief determinants of the optimal contract period are the serial correlations of customers' service valuations. If customers' valuations are invariant then permanent contracts suffice, but if valuations at any two dates are imperfectly correlated then a limited period may be efficient.

To take a simple example, let time be discrete and suppose that each customer's valuations follow the same finite-state ergodic Markov chain with a transition matrix M having H as its invariant distribution. Let $\alpha = (h_i w_i)$, where h is the stationary frequency vector corresponding to H and w_i is the service quality (assumed to be invariant) accorded customers initially in state i with service valuation v_i, and let $v = (v_i)$. Then the per-period average social cost of a contract of length T is

$$\mathcal{L}(T) = [c + \sum_{t=1}^{T} L(t)]/T,$$

where c is the average social cost per customer of recontracting and

$$L(T) = \alpha[I - M^{t-1}]v$$

measures the expected allocative efficiency loss in the t-th interval. Choosing the contract period T to minimize $\mathcal{L}(T)$ then provides an optimal tradeoff between contracting costs and allocative inefficiencies. The key point is that the length of the optimal contracting period is generally increasing in both the cost of recontracting and the serial correlation of customers' valuations.

This motive for forward contracting is closely related to the one revealed by Proposition 3. There it was shown that a few priority classes suffice to realize most of the efficiency gains. Thus, if each customer's valuations are highly serially correlated then a few contracts revised infrequently can cheaply supplant the continual variation in spot prices required by a spot market.

Uncertain Valuations

A deficiency in the schemes above is that they are sensitive to the assumption that each customer's private information, summarized in his type v, fully identifies his subsequent valuation of service. In practice, at the time of contracting a customer's information allows only a conditional probability distribution $G(\tilde{v}|v)$ for his valuation $\tilde{v}$ of service at the time of delivery.

This situation has been studied by Pitbladdo (1985) in the case that customers' valuations are statistically independent of each other and the supply, and the marginal distribution

$$G(\tilde{v}) = \int_0^1 G(\tilde{v} \mid v)\, dH(v),$$

is surely the realized distribution of valuations; i.e., there is no aggregate demand uncertainty.[8] Assume, moreover, that the seller can observe whether a customer demands service at the time of delivery, namely, whether his switch is ON. In this case, efficient schemes are analogous to the ones above: each customer selects a priority based on his current type. However, even if the customer's priority allows service in some subsequent contingency, he is served only if he demands service based on his realized valuation. Moreover, he pays the <u>contingent</u> price $\hat{p}$ assigned to his selected priority class only if he is served, which is the feature that affects his choice of whether to demand: he demands service only if $\tilde{v} \geq \hat{p}$. That is, the novel feature is that delivery is conditioned both on the customer's initially selected priority and his subsequent demand for service. Methods for computing the optimal contingent prices are included in Pitbladdo (1985) and Chao, Oren, Smith, and Wilson (1987, Section 5).

This is an optimal scheme given the prohibition of communication between the seller and the customers after the initial contracting, excepting only the observation of expressed demand. It illustrates most clearly, therefore, the inefficiencies attributable to imperfect communications. For example, if one customer selects a higher priority and therefore a higher contingent price than a second, then even though their realized valuations are in the same order, the first may not be served when the second is. This occurs whenever the one's valuation is less than his contingent price (so he does not demand) and the second's valuation is above

his contingent price, which in turn is high enough to merit service.

This scheme can be amended if there are multiple service times. Assume that a customer's valuations $\tilde{v}_t$ at successive dates t are serially correlated conditional on the initial type v. Then from the observation that a customer does not demand at the price $\hat{p}_1$ at time 1, the seller infers that $\tilde{v}_1 < \tilde{p}_1$, which makes it more likely that $\tilde{v}_2 < \hat{p}_2$ at time 2; similarly, the reverse is true if the customer does demand at time 1. Therefore, the optimal sequential scheme has the seller revising the customer's priorities and contingent prices for time 2 based on whether or not they demanded service at time 1.

The implication of these results is that in more general settings it is the contingent price $\hat{p}$ that provides the correct signal to customers. (We use the uncontingent price p in the present exposition only to simplify exposition.)

5. COMPETITIVE RATIONING

We examine next the incentives for profit-maximizing firms in competitive markets to offer priority service. Our focus is the inefficiency that accompanies imperfect competition among oligopolistic firms. We show that firms in an established oligopoly may not have incentives to differentiate their delivery conditions. Consequently, entry of additional firms can improve efficiency in some cases. This gain in allocative efficiency nevertheless sacrifices productive efficiency if there are gains from pooling supplies, as in the case that firms' supplies are imperfectly correlated.

We study differentiation and pricing in a symmetric oligopoly

comprising several identical firms. There are two basic cases. In one each firm offers a single class of service, and in the second one or more firms offer differentiated priority classes. Concise theoretical characterizations are elusive so numerical examples illustrate the main features of both cases.

Symmetric Firms with Single-Class Service

We describe briefly the formulation of a model of oligopoly among n symmetric firms, each offering a single class of service. The formulation anticipates the key feature that generically equilibria are asymmetric.[9]

Assume that each identical firm $i = 1,\ldots, n$ offers a single option $\langle w_i, p_i \rangle$ consisting of an average quality w_i and a price p_i. Each firm serves its customers in random order. Thus, a customer of type v selecting firm i obtains the expected gross benefit.

$$\bar{u}(v,q_i) = \int_0^{q_i} u(v,\ w(r))\ dr/q_i\ ,$$

if in total q_i customers elect service from firm i. Anticipating an asymmetric equilibrium, assume that q_i is decreasing and p_i is increasing in the index i of the firm. Then firm i is preferred by an interval $[v_{i-1},\ v_i]$ of customer types, where $v_n = 1$,

$$\bar{u}(v_i, q_i) - p_i = \bar{u}(v_i, q_{i+1}) - p_{i+1}$$

if $i < n$, and $\bar{u}\ (v_0,\ q_1) = p_1$ or $v_0 = 0$. The number of customers served by firm i is $q_i = H(v_i) - H(v_{i-1})$, and firm i obtains the profit $p_i q_i$.

In the game that we consider, the firms first choose their prices

simultaneously. Knowing these prices and anticipating the demands or qualities at the firms, each of the customers then (simultaneously) selects one firm, if any. Thus, in a Nash equilibrium, each firm i selects its price p_i to maximize its profits $p_i q_i$, anticipating other firms' prices and how its demand q_i depends on its price. The conditions that characterize an equilibrium are sufficiently complicated to omit them here. However, they can be solved numerically for particular examples.

• Table 2 illustrates the standard example modified to specify that each firm's supply is uniformly distributed on the interval [0, 1/n]. Note that the allocation of customers to firms becomes increasingly efficient as the number of firms increases.

Table 2
Equilibrium Among Competing Firms
Capacity Divided Equally Among Firms

Standard Example: $u(v,w) = vw$, $H(v) = v$, $w(r) = 1 - r$, $F_i(s) = ns$

Firms n	Class i	Value v_{i-1}	Qual w_i	Price $\hat{p}_i$	Quant q_i	ConSur CS_i	Profit Π_i	TotSur TS_i
2	2	.736	.736	.445	.264	.082	.086	.169
	1	.424	.688	.424	.312	.033	.091	.124
	Totals:				.576	.116	.177	.293
4	4	.865	.729	.410	.135	.052	.041	.092
	3	.717	.706	.395	.147	.041	.041	.082
	2	.554	.674	.380	.163	.028	.042	.070
	1	.366	.623	.366	.188	.011	.043	.054
	Totals:				.634	.132	.166	.298
8	Totals:				.663	.139	.160	.299
12	Totals:				.673	142	.158	.300
∞	Totals:				.6936	.1466	.1532	.2998

Reitman (1986) provides a general proof that allocative efficiency is achieved in the limit as the number of firms increases. The model assumes that industry capacity is 'divided equally' among identical firms.

Specifically, assume that with a finite number n of firms, each firm's technology of supply is specified by the quality function $w_n(r)=w(nr)$, as in the example above. It follows that if $\bar{u}_n(v,q)$ is a customer's expected gross benefit at one of n firms having q customers, then $\bar{u}_n(v,q) = \bar{u}_1(v,nq) \equiv \bar{u}(v,nq)$. Use $t = i/n$ to indicate a firm's fractile rank in the distribution of prices offered. Then $p(t) = p_i$ denotes the i-th firm's price, and $v(t) = v_i$ is the highest type selecting that firm. $Q(t)= \sum_{j \le i} q_j$ specifies the number of customers served by the i firms with the lowest prices and highest demands, and $\Pi(t) = \sum_{j \le i} p_j q_j$ is the profit of these firms. Reitman shows that the limiting distribution Q(t) obtained as $n \to \infty$, assuming it is twice differentiable, is an efficient allocation of customers among firms. At the limit, the allocation is characterized by conditions on the price distribution that summarize customers' self-selection and firms' profit maximization:

$$p'(t) = \bar{u}_q(v(t),Q'(t))Q''(t),$$
$$p(t) = -\bar{u}_q(v(t),Q'(t))Q'(t).$$

These imply that all firms obtain the same profit $\Pi^\circ \equiv \Pi'(t) = p(t)Q'(t)$.[10]

- In the standard example adapted to one version of Example 2, $u(v,w) = vw$, $H(v) = v$, and $w(r) = -r$. In this version, $\bar{u}(v,q) = v\bar{w}(q)$ where $\bar{w}(q) = -q/2$ is a customer's expected waiting time in the queue of length $q = Q'(t)$ at firm t. The asymptotic equilibrium has $v(t) = Q(t) = t^{2/3}$, $p(t) = \frac{1}{3} t^{1/3}$, and $\Pi^\circ = 2/9$. The aggregate of the firms' profits is 2/9, and also this is the aggregate of customers' waiting costs.

- In the version adapted to example 1, $w(r) = 1 - r$ for $r \le 1$, and

therefore $\bar{w}(q) = 1 - q/2$ if $q \leq 1$ and $\bar{w}(q) = 1/2q$ if $q > 1$. In the asymptotic equilibrium, the lowest type $\underline{v} \approx .3064$ served is the real root of the equation $\underline{v}[3 + 2v]^2 = 4$. The allocation and prices are

$$v(t) = Q(t) + \underline{v} = \underline{v}^{1/3} \; [\underline{v} + 3t/2]^{2/3},$$

$$p(t) = \frac{1}{2} \underline{v}^{2/3} \; [\underline{v} + 3t/2]^{1/3},$$

and the resulting profit and consumer surplus are

$$\Pi^{\circ} = \frac{1}{2} \underline{v} \qquad \text{and} \qquad CS^{\circ} = \frac{1}{2}[1 - 2\underline{v} - \underline{v}^2].$$

These are the results tabulated in the last row of Table 2. Prices increase from $p(0) = \frac{1}{2} \underline{v} \approx .1532$ to $p(1) = \frac{1}{2} \sqrt{\underline{v}} \approx .2768$ while the density of customers per firm declines from $Q'(0) = 1$ to $Q'(1) = \sqrt{\underline{v}} \approx .5536$.

A limitation of these results is that they establish only the allocative efficiency of the assignment of customers to firms. The productive efficiency of the dispersal of supplies among firms is not addressed. The culprit is the assumption that capacity is 'divided equally' among firms. In Examples 1 and 2, one interpretation is that in each contingency the industry supply $s = \sum_i s_i$ is divided equally: if there are n firms then firm i has available the supply $s_i = s/n$. In Example 2 dividing supply equally is a natural approximation. In Example 1 where supplies are uncertain, however, it implies that firms' supplies are perfectly correlated. When firms' supplies are perfectly correlated it is efficient to assign greater supplies to firms serving higher types. As the number of firms increases, the productive inefficiency of equal division increases and it diminishes the surplus that can be realized.

The alternative interpretation is that firms' supplies are imperfectly correlated: although capacities are equal, supplies are not. For instance, if firms' supplies are statistically independent and each firm has the same

distribution $F_i(s_i) = ns_i$ as in Table 2, then with many firms it is almost certain that the total supply is precisely $s = .5$. In this case, it is efficient to serve only those types $v \geq \underline{v} = .5$ and realize the greater surplus .375. This productive inefficiency could be eliminated by pooling the firms' supplies to eliminate most of the uncertainty that individual firms encounter: a spot market for wholesale trades is one means to accomplish this.[11]

• To illustrate, consider the standard example modified so that the distribution F(s) of the aggregate supply s is the Normal distribution with means $\mu = 1$ and standard deviation $\sigma = .5$. Consider three cases:

(a) If firms' supplies are independently and identically distributed, then each firm's supply has the Normal distribution with mean μ/n and standard deviation $\sigma/\sqrt{n}$.

(b) If each firm obtains the supply s/n so that firms' supplies are perfectly correlated, then each firm's supply has the Normal distribution with mean μ/n and standard deviation σ/n.

(c) Alternatively, the firms can pool their supplies and in each contingency allocate the available aggregate supply among themselves in priority order: the firms serving higher types have higher priority. Each firm receives some supply, if any, for its customers only after firms with higher priorities have served their customers.

To facilitate comparisons, all three cases assume that the firms price their services non-cooperatively, although this is least plausible in case (c). Table 3 summarizes the aggregate results for several numbers of firms.[12] If $n = 1$ then all three cases yield the same consumers' surplus, profits, and total surplus: .106, .232, and .338 respectively. Note that in case (a) the total surplus declines steadily as the number of firms increases above $n = 2$; this reflects the foregone benefits of pooling supplies. The differences in the table show that the organization of supply relationships among firms has significant effects on productive efficiency.

Thus, allocative efficiency is a limited benefit of competition among firms when account is taken that overall efficiency is sensitive to the productive role of cooperative relationships among firms.

Competitive Differentiation of Priorities

We now examine firms' incentives to differentiate their service conditions. Greater differentiation of priority service creates efficiency gains that might be captured in part by the innovative firm. These efficiency gains arise both from the finer sorting of customers, and from the greater market penetration that ensues. This is generally the case for a monopoly, as evident in Table 1 for instance, and therefore multiplicity of priority classes is predictably the norm for a monopolist. We argue nevertheless that firms' differentiation of priority services can not be predicted generally. For this purpose we study a symmetric duopoly that is not subject to the threat of entry.

Table 3
Comparison Among Three Productive Regimes
$u(v,w)=vw$, $H(v)=v$, $w(r)=1-r$, $F(s) \sim N(\mu,\sigma)$, $\mu=1.0$, $\sigma=0.5$

		n=2			n=3	
Case	ConSur	Pft	TotSur	ConSur	Pft	TotSur
(a)	.233	.146	.378	.246	.117	.362
(b)	.251	.161	.412	.271	.145	.416
(c)	.285	.145	.430	.333	.105	.438
		n=4			n=∞	
Case	ConSur	Pft	TotSur	ConSur	Pft	TotSur
(a)	.251	.100	.351	.250	.000	.250
(b)	.281	.136	.417	.302	.117	.419
(c)	.353	.087	.439	.362	.080	.442

To examine a firm's incentive to offer more than one class of service, we use numerical illustrations based on the standard example. Recall that preferences are given by $u(v,w) = vw$, where types have the

distribution $H(v) = v$, and the quality from service order r is $w(r) = \bar{F}_i(r)$ if F_i is the distribution function of the firm's supply s_i. Random rationing within a service class that has q customers yields, therefore, the expected gross benefit $v\bar{w}(q)$, where $\bar{w}(q) = \int_0^q \bar{F}_i(Q+r)\, dr/q$ and Q is the number of customers served in any higher priority classes offered by the firm. The illustrations assume that F_i is the normal distribution with mean $\mu_i = 1/n = .5$ and standard deviation $\sigma/\sqrt{n} = .5/\sqrt{2}$. If the firms' supplies are independent, then the distribution F of total supplies agrees with the one in Table 3, namely the mean is $\mu = 1$, and the standard deviation is $\sigma = .5$.

• Table 4 shows the numerical results for an example in which firm 1 offers a single class of service and firm 2 offers one or more priority classes. Firm 2 is the one that has higher prices, less demand, and lower profit when both firms offer single-class service. In all cases, the salient feature of the equilibria is that if one firm, here firm 2, offers more than one priority class then there is a single equilibrium in which, from customers' perspective, the other firm's service offers the second-highest quality.[13] Both firms' profits decline if one offers more than a single class of service; in particular, the innovating firm is unable to capture a share of the increment in surplus that its differentiation produces.

Table 4
Duopoly with One Firm Offering Multiple Classes
$F_i(s) \sim N(1.0/2, 0.5/\sqrt{2})$

Number of Classes of Firm 2	Profits			Surplus	
	Firm 2	Firm 1	Total	Consumers	Total
1	0.071	0.075	0.146	0.232	0.378
2	0.069	0.060	0.129	0.262	0.391
3	0.066	0.057	0.123	0.269	0.393
4	0.065	0.056	0.122	0.271	0.393

Similar conclusions prevail in all numerical examples I have studied, including for instance the standard example in which $F_i(s) = 2s$. Neither firm has an incentive to increase its offering to more than one class of service, and the advantages of simultaneous differentiation by both firms are unclear due to the nonexistence of equilibria.[14]

• Further evidence appears in examples in which the two firms pool their supplies and serve all classes in priority order. Thus, for each firm its available supply to serve a specified class is the residual after all classes of both firms with higher priority have been served from their total supply. This evidently requires cooperation between the firms, but it is an extreme assumption that tests the limits of the firms' abilities to claim a share of the incremental surplus from differentiation. To exclude monopoly pricing by a cartel, we continue to suppose that the firms choose their prices non-cooperatively for the priority classes they offer. In this example, equilibria exist for each possible priority ordering of the firms' service classes. In Table 5 we show the aggregate results for equilibria of examples with two, three, and four classes; all others are obtained by interchanging the roles of the firms.

The examples in Table 5 provide little encouragement that

differentiation of priority services will occur non-cooperatively. Total profits evidently decline with further differentiation, including cases not shown here. Either firm could gain slightly by offering the highest classes while the other offers the low class, but at considerable expense to the other firm; while for firm 1, offering a second lower class is unprofitable.[15] Cooperative sharing of supplies could be accompanied by compensatory transfers, but if so then the dominant motive is to sustain the largest aggregate profits via a single class for each firm. It appears that only a monopoly cartel that pools supplies and coordinates prices is assured to differentiate. Of course these conclusions are not meant to apply to public utilities, since ordinarily they serve non-overlapping districts.

Table 5
Competition Between 2 Firms with Multiple Classes
Pooled Supply, F(s) ~ N (1.0, 0.5)

Classes Offered		Profit			Surplus	
Firm 2	Firm 1	Firm 2	Firm 1	Total	Cons.	Total
High	Low	0.104	0.041	0.145	0.285	0.430
High, Middle	Low	0.106	0.032	0.138	0.296	0.434
High, Low	Middle	0.087	0.038	0.125	0.309	0.434
Middle, Low	High	0.036	0.093	0.128	0.307	0.435
Top, High	Middle, Low	0.096	0.025	0.120	0.318	0.438
Top, Middle	High, Low	0.071	0.026	0.097	0.342	0.439
Top, Low	High, Middle	0.071	0.027	0.098	0.341	0.439

My conclusion from these and other numerical examples is that it is not possible, based on the formulation used here, to construct a theory that would predict that firms in a non-cooperative oligopoly have an incentive to differentiate their service classes. Nonexistence of equilibria presents one evident difficulty, but more importantly, those equilibria that do exist show that firms have an incentive not to differentiate. It seems clear,

therefore, that in competitive markets the efficiency gains obtained from differentiation of priority services must depend on entry of additional firms, or at least the threat of entry.

6. CONCLUSION

We have examined the rationing of scarce supplies via priority service in state enterprises and among firms in competitive markets. The gains from efficient rationing, as compared to random rationing, stem primarily from the diversity of customers' preferences. When these gains can not be obtained directly via spot markets, they can be realized in part by contingent forward contracts that specify the customer's priority or service order. Only a few priority classes suffice to realize most of the efficiency gains. The relative pricing of these contracts is determined by the need to induce customers to self-select their optimal contracts based on their preferences. In the simplest cases, the price can be interpreted as the expected spot price for equivalent service. Priority service is Pareto superior to random rationing with only the simplest 'equal dividends' rule for redistribution of revenues. A state enterprise or public utility has a variety of schemes available to implement priority service, some of which rely entirely on customers to assess and provide information. The information revealed by customers' responses to priority service is also sufficient for capacity planning. Except for a monopolist, profit-maximizing firms lack clear incentives to differentiate priority services, except possibly under threat of entry. Competitive markets, divided into many small firms offering single-class service, simulate priority service via dispersed prices and therefore dispersed qualities. These markets

approximate an efficient allocation of dispersed supplies; hence, entry is apparently an important factor in achieving efficiency via market forces. However, these gains must be compared with those obtainable from pooling supplies. If pooling is important then a regulated public utility may be superior, since oligopolistic firms engaged in pooling arrangements may also lack strong incentives to differentiate their services.

The theory of priority service is similar to standard theories of product differentiation in many respects. However, a novel feature is that qualities are affected endogenously by customers' selections. This accounts for the major differences in the results, such as the disincentives for oligopolistic firms to offer multiple service classes.

The immediate practical applications of priority service are in the capital-intensive industries subject to peak loads. We emphasize electric power as an important context in which priority service can improve allocative efficiency. It can also substitute for capacity expansion, or it can provide evidence from customers' market behavior about their willingness to pay for improved reliability.

ENDNOTES

1. By 'high' priority we mean the front of the queue for service; unfortunately, English uses 'high' to mean a low numbered position in the queue.

2. Priority service contracts from the viewpoint of customers can also be interpreted as option contracts from the viewpoint of the enterprise. That is, they allow the enterprise to 'call' units of supply from customers. However, because they differ in significant ways from the put and call options traded in securities markets, we avoid this terminology.

3. The focus here on pecuniary externalities, as opposed to externalities from crowding that directly diminish the quality of amenities, distinguishes priority service from the topics addressed in the theory of clubs and local public goods; cf. Scotchmer (1985).

4. A more general model that represents a customer's demand as a load-duration profile is analyzed by Chao, Oren, Smith, and Wilson (1986a).

5. All numerical examples are computed by a program PRISM written in the STSC version of the APL language, available on request from the author.

6. This conclusion is altered materially if there is demand uncertainty. See Chao and Wilson (1987a) for an extension to this case. The key feature is that the expected spot revenue differs from the revenue collected by charging the expected spot price if total demand is correlated with the 'marginal' demand that determines the spot price.

7. This formula can be altered to include a factor representing the probability that the customer demand service, as in the case of electric power where the customer's switch is typically ON only a portion of the time. Wilson (1987) addresses the case that customers' ON rates are correlated.

8. Recall that throughout we allow aggregate demand to be uncertain in the sense of being contingent on publicly observable variables, such as temperature, which merely condition the price associated with each priority. Also, if aggregate demand is an independently random, invertible function of a base demand, then this randomness can be incorporated into the supply distribution. Some kinds of correlations among customers' valuations or their ON rates allow similar treatments. Here, as elsewhere, these types of aggregate demand uncertainty cause no difficulty in the theoretical development.

9. Generic asymmetry of the equilibria is proved by Reitman (1986) for n > 2. The special case in Example 2 that n = 2 and w(r) = - r/s has a symmetric equilibrium.

10. Firms can also differentiate their services further by investments in capacity. Reitman (1986) shows in the context of Example 2 that if firms' capacity costs are linear then firms serving higher types choose greater capacities, but the net effect is small. The allocative efficiency of the asymptotic equilibrium remains valid: firms choose equal capacities and obtain zero profits net of capacity costs.

11. In a third interpretation, each firm's supply distribution does not depend on the number of firms: $w_n(r) = w(r)$, reflecting entry of firms with identical supply technologies. The trivial result is that with many firms, prices and profits are zero and the full potential consumers' surplus is realized (0.5 in the context of the example used in Table 2).

12. In case (b), $n = \infty$, the results reported are actually for n=16.

13. As in all the examples I have studied, each other ordering produces an apparently unique solution to the necessary conditions for a solution, but fails to be an equilibrium because the prices and qualities do not match the ordering of market segments served. It appears that no asymmetric equilibrium exists when both firms offer two or more classes. All such conclusions are subject, of course, to the fallibility of numerical methods.

14. These conclusions are reinforced by a general result of Reitman (1986): If each of two identical firms offers a continuum of service orders then, subject to a proviso, there is no pure-strategy equilibria (symmetric or asymmetric) of the game in which each chooses its price schedule. The proviso is an assumption that given firms' price schedules, including disequilibrium ones, customers allocate themselves efficiently, possibly including side payments among themselves.

15. Firm 2's profit declines slightly below 0.106 if it offers three high classes and firm 2 offers the low class for a profit of 0.031; total profits decline to 0.156.

REFERENCES

[1] Bohn, Roger; M. Caramanis; and Fred Schweppe (1984): "Optimal Pricing in Electrical Networks over Space and Time." Rand Journal of Economics, 15, 3690-376.

[2] Boiteux, M. (1960): "Peak Load Pricing," Journal of Business, 33, 157-179.

[3] Caramanis, M.; Roger Bohn; and Fred Schweppe (1982); "Optimal Spot Pricing: Practices and Theory," IEEE Transactions on Power Apparatus and Systems, PAS-101, 3234-3245.

[4] Chao, Hung-po: Shmuel Oren; Stephen Smith; and Robert Wilson (1986a): "Multi-level Demand Subscription Pricing for Electrical Power," Energy Economics, 4, 199-2178.

[5] Chao, Hung-po; Shmuel Oren; Stephen Smith; and Robert Wilson (1986b): Priority Service: Unbundling the Quality Attributes of Electric Power, Palo Alto, CA: Electric Power Research Institute, EW 4851-November.

[6] Chao, Hung-po Shmuel Oren; Stephen Smith; and Robert Wilson (1987): Selected Papers on Priority Service, Palo Alto, CA: Electric Power Research Institute, P-5350, August.

[7] Chao, Hung-po, and Robert Wilson (1987a): "Priority Service: Pricing, Investment, and Market Organization," American Economic Review, to appear.

[8] Chao, Hung-po, and Robert Wilson (1987b): "Optimal Contract Periods for Priority Service." Palo Alto, CA: Electric Power Research Institute.

[9] Harris, Milton; and Artur Raviv (1981): "A Theory of Monopoly Pricing Schemes with Demand Uncertainty," American Economic Review, 71, 347- 365.

[10] Marchand, M. (1974): "Pricing Power Supplied on an Interruptible Basis," European Economic Review, 5, 253-274.

[11] Mendelson, Haim; and Seungjin Whang (1976): "Optimal Incentive-Compatible Priority Pricing for the M/M/1 Queue," Rochester, NY: Graduate School of Management, University of Rochester, October.

[12] Mookherjee, Dilip (1986): "Involuntary Unemployment and Worker Self-Selection," Stanford Business School.

[13] Oren, Shmuel (1987): "On the Existence of Price Functions for Two-Dimensional Menus," Palo Alto, CA: Electric Power Research Institute, July.

[14] Pitbladdo, Richard (1985): "Interruptible Service Under Individual Uncertainty," Stanford University, Mimeo, November. See also Imperfect Communication in the Delivery of a Nonstorable Good, Ph.D. Dissertation, August 1986.

[15] Reitman, David (1985): "Pricing, Quality, and Priority Service in Congested Markets," Stanford University, mimeo, October.

[16] Reitman, David (1986): Competition in Congested Markets, Ph.D. dissertation, Stanford University, September. See also "Price Dispersion and Quality Differentiation in Congested Markets," February 1987, and "Competitive Priority Service Mechanisms," May 1987, Ohio State University,.

[17] Scotchmer, Suzanne (1985): "Two-Tier Pricing of Shared Facilities in a Free-Entry Equilibrium," Rand Journal of Economics, 16, 452-472.

[18] Telson, M.L. (1975): "The Economics of Alternative Levels of Reliability for Electric Power Generation Systems," Bell Journal of Economics, 6,679-694.

[19] Vickrey, William (1971): "Responsive Pricing of Public Utility Services," Bell Journal of Economics, 2, 337-346.

[20] Wilson, Robert (1987): "Service Reliabilities Corrected for ON Rates," Palo Alto, CA: Electric Power Research Institute, February.